



저작자표시-비영리-변경금지 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.

다음과 같은 조건을 따라야 합니다:



저작자표시. 귀하는 원저작자를 표시하여야 합니다.



비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.



변경금지. 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 [이용허락규약\(Legal Code\)](#)을 이해하기 쉽게 요약한 것입니다.

[Disclaimer](#)

보건학석사 학위논문

Comparison of DNA Methylation Profiles in Sibships Discordant for Intrauterine Exposure to Maternal Gestational Diabetes Mellitus

어머니의 임신성 당뇨 진단 여부에 따른 자녀들의
DNA 메틸화 수준 비교: 형제자매 내의 비교 연구

2015 년 8 월

서울대학교 보건대학원
보건학과 유전체역학 전공
김 은 애

Comparison of DNA Methylation Profiles in Sibships Discordant for Intrauterine Exposure to Maternal Gestational Diabetes Mellitus

어머니의 임신성 당뇨 진단 여부에 따른 자녀들의 DNA 메틸화 수준 비교: 형제자매 내의 비교 연구

지도교수 성주헌

이 논문을 보건학 석사학위논문으로 제출함

2015년 5월

서울대학교 보건대학원

보건학과 유전체역학 전공

김 은 애

김은애의 석사학위논문을 인준함

2015년 7월

위 원 장	<u>원 성 호 (인)</u>
부 위 원 장	<u>정 해 원 (인)</u>
위 원	<u>성 주 헌 (인)</u>

Abstract

Introduction: *In utero* diabetic exposure is reported to be associated with predisposition to metabolic diseases in later life, supporting the hypothesis of Developmental Origins of Health and Disease (DOHaD). One of the underlying mechanisms proposed so far is by alterations in DNA methylation levels, which are associated with gene expression. To assess epigenetic modifications at DNA methylation levels associated with prenatal exposure to diabetes, we conducted a discordant sibship study, in which a total of 38 siblings showing discordance for intrauterine exposure to maternal gestational diabetes mellitus were recruited.

Methods: We collected data on birth outcomes and data of anthropometric, physiological, and biochemical measurements after recruitment. DNA methylation levels of peripheral leukocytes were examined at over 485,000 CpG sites using Illumina Infinium HumanMethylation 450 BeadChip assays. To obtain an overview of methylation profiles across samples, intersample distance was measured using unsupervised hierarchical clustering methods. Within-pair differential methylation analysis was performed on genomic region levels including CpG islands, genes, and promoters, as well as on CpG site levels. Pathway and function analysis was conducted to identify biological implications associated with differentially methylated sites and regions.

Results: In a Manhattan distance-based unsupervised hierarchical clustering analysis of methylation levels across samples, three of the sibling pairs were closest to each other, while most of the rest were closely related to each other

based on prenatal GDM exposure status, independent of siblings. A within-pair differential analysis on CpG site levels showed that 18 sites were differentially methylated at p -values $< 10^{-5}$. Among those significantly differential CpG sites was cg08407434, which is associated with HNF4A loci, with the mean pairwise difference of methylation levels of 1.3% (p -value = 9.1×10^{-7}). In the region-level differential methylated analysis, 23 genes and 24 promoter regions were differentially methylated at p -value $< 10^{-3}$. In the pathway analysis, the main pathway overrepresented by hypermethylated gene regions was immune responses.

Discussion: To the best of our knowledge, this is the first epigenome-wide study investigating methylation profiles in discordant sibships for their maternal GDM. Our findings suggest that *in utero* exposure to diabetic environment has epigenetic effects, particularly altering DNA methylation levels, and thus reinforce the evidence for the hypothesis of DOHaD.

Keywords: Gestational diabetes mellitus, DNA methylation, Epigenetics, Intrauterine environment, Developmental Origins of Health and Disease, Sibship study

Student Number: 2013-23582

CONTENTS

Abstract	i
List of tables	iv
List of figures	v
I. Introduction	1
1. Background	1
2. Objective	4
II. Method	5
1. Study population	5
2. Data collection	5
3. Descriptive statistical analysis	6
4. DNA methylation analysis	6
4.1. Genome-wide DNA methylation profiling	7
4.2. Unsupervised clustering analysis of DNA methylation	9
4.3. Differential methylation analysis	9
5. Function and pathway analysis	11
III. Results	12
1. Study population	12
2. DNA methylation analysis	17
2.1. Exploratory analysis	17
2.2. Differential methylation analysis	21
3. Function and pathway analysis	38
IV. Discussion	41
V. References	44
VI. Abstract in Korean (국문 초록)	48

List of tables

Table 1. Descriptive characteristics of the study population.	14
Table 2. Summary of DNA methylation data by site/region.	17
Table 3. 50 top-ranked differentially methylated CpG sites.	24
Table 4. 50 top-ranked differentially methylated CpG island regions.	29
Table 5. 50 top-ranked differentially methylated gene regions.	32
Table 6. 50 top-ranked differentially methylated promoter regions.	35

List of figures

Figure 1. Comparison of the density distributions of raw and preprocessed β values for the Infinium 1 and Infinium 2 probes.	8
Figure 2. Histogram of β value frequency by the sample group.	18
Figure 3. Hierarchical clustering of β values.	20
Figure 4. Multidimensional scaling plot of the β values.	21
Figure 5. Volcano plots comparing methylation within siblings on site levels.	23
Figure 6. Comparison of differences of β values within sibling pairs.	27
Figure 7. Volcano plots comparing methylation within siblings on region levels.	28
Figure 8. Pathways associated with differential methylated regions.	39
Figure 9. Diseases and biological functions associated with differential methylated regions.	40

I. INTRODUCTION

1. Background

Gestational diabetes mellitus (GDM) is a condition of glucose intolerance, first recognized and diagnosed during pregnancy. The prevalence of GDM in advanced economies varies between populations, ranging 1.7-11.6% [1]. Approximately 2-5% of Korean pregnant women are affected by GDM [2]. Increased levels of pregnancy-related hormones, such as lactogen, estrogen, and prolactin, induce glucose resistance, resulting in elevated blood sugar levels. GDM usually resolves after pregnancy; however, the condition poses a risk of adverse pregnancy complications, such as congenital malformations, increased birth weight and adiposity, and perinatal death [3, 4]. Not only are women with GDM at an increased risk of type 2 diabetes mellitus (T2D) after pregnancy, but their offspring are also at an increased risk of metabolic diseases in their later life, such as higher body mass index (BMI), higher systolic blood pressures (SBP), greater adiposity, impaired glucose tolerance (IGT) and defective insulin secretory response [4, 5]. An epidemiological study of sibling pairs discordant for maternal diabetes in the Pima Indian population showed a significant increase in BMI of 2.6 kg/m² for offspring born in diabetic than non-diabetic pregnancies [5].

Although the underlying mechanism of how GDM predisposes offspring to metabolic diseases in their adult life is still under discussion, the hypothesis of “Developmental Origins of Health and Disease (DOHaD)” is widely accepted [6]. In the context of DOHaD, an event occurring as early as prenatal

environmental exposures can program epigenetic modifications resulting in an elevated risk of chronic diseases in later life [6]. Such a concept has been supported by mounting evidence [7-9], including epidemiology studies of those exposed to different early-life environments. An increasing number of studies have assessed epigenetic modifications associated with diseases or phenotypes of interest using biomarkers, such as DNA methylation and histone modifications.

DNA methylation is by far the best studied epigenetic biomarkers to date. The most common form of DNA methylation in vertebrates is 5-methylcytosine (5mC), in which cytosine nucleotide in DNA is modified by the addition of a methyl (CH₃) group to its 5th carbon. It occurs almost exclusively at Cytosine-phosphate-guanine (CpG) dinucleotides, where cytosine and guanine are separated by one phosphate. The methylation at CpG sites at the promoters of genes is often associated with silencing of gene expression [10]. DNA methylation patterns are maintained through cell division by the action of DNA methyltransferase 1 (DNMT1). Profiling genome-wide DNA methylation levels across CpG sites has been realized by the rapid development of laboratory techniques, including the Illumina HumanMethylation450 Beadchip, which is one of the most popular arrays.

Previous studies reported that genes and pathways associated with differential methylation in relation to maternal GDM are involved in metabolic functions. Decreased methylation levels were found at the mesoderm-specific transcript (MEST) imprinted gene in cord blood and placenta of newborns born to mothers with GDM [11]. And also for obese adults, significant demethylation

at MEST was observed, suggesting association between epigenetic malprogramming and predisposition to obesity throughout life [11]. A pathway analysis showed that genes associated with potentially differentially methylated regions in the GDM-exposed group in comparison with the unexposed group were involved in the metabolic diseases pathway [12]. DNA methylation studies of offspring of diabetic mothers and offspring of non-diabetic mothers in Pima Indians found that genes with differentially methylated promoters were involved in the pathways of maturity onset diabetes of the young (MODY), type 2 diabetes, and Notch signaling [12]. Placental lipoprotein lipase (LPL) DNA methylation is associated with maternal GDM and maternal HDL-c levels [13]. Lesseur et al. found that infants exposed to GDM and those exposed to maternal prepregnancy obesity had higher placental leptin (LEP) methylation and that GDM serves as a mediator of the association between prepregnancy obesity and placental LEP DNA methylation [14].

Many epigenetic studies employed a population-based, case-control approach. Even though such studies allow for large-scale studies, they are susceptible to bias due to potential confounding factors, including genetic factors, age, sex, maternal effects, cohort effects, and intrauterine and environmental factors. As monozygotic (MZ) twins share genes and many environmental factors, the study design of disease-discordant MZ twins has thus been successful in controlling for such potential confounders and also allowed for pairwise comparison based on the phenotype of interest [15]. A discordant sibship design, on the other hand, is less able to control for potential confounding factors compared with MZ twin studies; however, differential methylation analysis within sibling pairs discordant for intrauterine

environment may be able to identify differential methylation regions associated with early-life environmental exposures, whilst modestly controlling for genetic and environmental confounding factors.

2. Objective

This study aims to examine whether there is evidence of differential methylation associated with intrauterine exposure status to diabetic environment. Specific aims of the study are (1) to detect differentially methylated CpG sites within sibling pairs, (2) to detect differentially methylated regions within sibling pairs, and (3) to identify biological pathways and functions overrepresented by differentially methylated regions.

II. METHODS

1. Study population

38 children born to 18 mothers having experienced both GDM and non-GDM pregnancies were recruited at Seoul National University Bundang Hospital for the study. Their ages at the recruitment ranged from 4 to 14 years old. Of the 38 offspring, 19 were exposed to maternal GDM, and the rest 19 were unexposed controls. Of the 18 sets of offspring, 16 were discordant full-sibling pairs for their maternal GDM, and the rest two were composed of three full-siblings with at least one sibling born after diagnosis of maternal GDM. In each set of three siblings, the oldest was born without having been exposed to maternal GDM, and the youngest exposed to maternal diabetes; the second child from one of the sibling sets was born to the mother with GDM, while the second child from another set were born to the mother diagnosed with IGT. This study was approved by The Seoul National University Hospital ethics Committee. Written informed parental consent was obtained before the inclusion in the study.

2. Data collection

For diagnosis of GDM for mothers, a two-step approach was employed: (1) 50-g oral glucose tolerance test (OGTT) was initially used for screening at-risk women of GDM. (2) For those who tested positive at 50-g OGTT (plasma or serum glucose concentration 1 hour after a 50-g OGTT was higher than 140

mg/dL), 100g OGTT was performed at 28 ± 2 weeks. If at least one of the following criteria was met (≥ 96 , 180, 155, and 140 mg/dL for fasting, one-hour, two-hour, and three-hour plasma glucose concentration, respectively), the diagnosis of GDM was made. Available data of offspring on birth outcomes were collected. Data for anthropometric measurements performed at their visit to Seoul National University Bundang Hospital after the recruitment were collected. Additionally, their blood samples were extracted for physiological and biochemical profiling and for DNA methylation analysis.

3. Descriptive statistical analysis

For sample description by group, mean \pm standard deviation was presented for continuous variables and counts were presented for categorical variables. To compare continuous variables between groups while reducing confounding effects due to sibling correlations, we fit a linear mixed model, where each measurement is a response variable, grouping is adjusted as fixed effects, and siblings were added as random effects. *P*-value for fixed effects of the group was computed using Satterthwaite's approximations, after sibling random effects were adjusted for. Linear mixed modelling was performed using lmerTest package [16] in R (v. 3.1.1) [17]. Pearson's Chi-square test was used to compute *p*-values using R [17].

4. DNA methylation analysis

4.1. Genome-wide DNA methylation profiling

DNA was extracted from the peripheral leucocyte samples. The DNA samples were treated with sodium bisulfite and applied to the Illumina Infinium HumanMethylation 450 BeadChip assays, following the standard manufacturer's protocols. The Infinium platform assayed > 485,000 CpG loci to quantify DNA methylation status for each site.

Raw methylation data in IDAT formats were load into the R [17] environment and processed using RnBeads, an R Bioconductor package which allows for comprehensive analysis of DNA methylation data [18]. The raw intensity values were read and converted into beta (β) values of individual CpG sites. β value is defined as the proportion of the total intensity coming from the methylated channel, as follows,

$$\beta = \frac{Meth}{Meth + Unmeth + \alpha}$$

where *Meth* and *Unmeth* are intensities measured by methylated probes and unmethylated probes, respectively, and α , usually set to 100, is a constant for regularizing β when both *Meth* and *Unmeth* are small.

For 485,577 probes, a series of probe filtering was conducted to remove (1) SNP-enriched probes (n = 4,713), (2) probes with detection *p*-value ≥ 0.05 (n = 821), (3) probes outside of CpG context (n = 3,152) and (4) probes on sex chromosomes (n = 11,193). A total of 465,698 probes were included for analysis. Technical differences between two different designs of the Illumina 450k Methylation Array, Infinium 1 and Infinium 2, were reduced using beta-

mixture quantile normalization (BMIQ) [19]. The method adjusts the probe distribution of Infinium 2 based on that of Infinium 1 in order to make their statistical distributions comparable. In addition, a background subtraction method, NOOB, was applied to reduce technical variation in background fluorescence signal [20]. The distribution of the β values of the raw and preprocessed data by array design is compared in Figure 1.

Genomic regions associated with one or more CpG sites, including genes, promoters, and CpG island regions, were inferred. For annotations of genes and promoters, Ensemble gene definitions were downloaded using the biomaRt package [21, 22]. Annotations for CpG islands were obtained using the CpG island track from UCSC Genome browser [23]. Gene-related regions were defined as the whole locus from transcription start site (TSS) to transcription end site (TES). Promoters were defined as the region spanning 1,500 bases upstream and 500 bases downstream of the TSS of the corresponding gene.

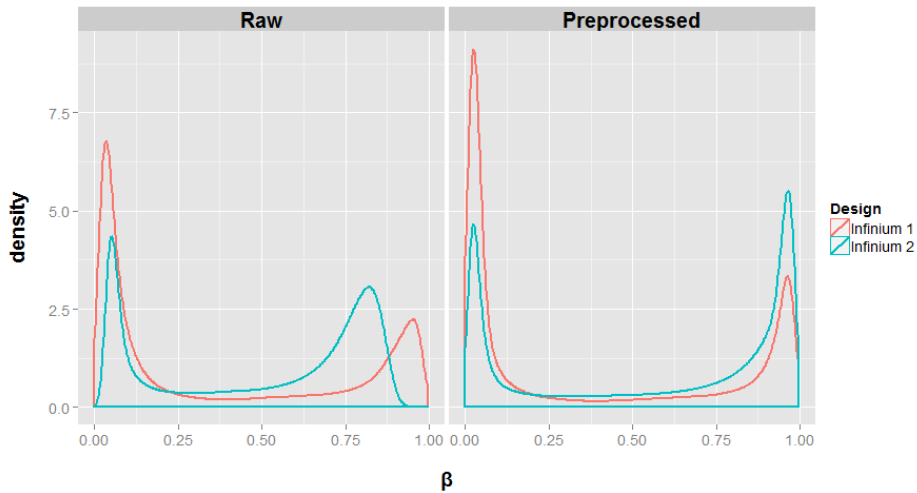


Figure 1. Comparison of the density distributions of raw and preprocessed β values for the Infinium 1 and Infinium 2 probes.

4.2. Unsupervised clustering analysis of DNA methylation

Unsupervised learning techniques were used for sample clustering, in order to inspect the data set for a signal in the methylation values associated with samples' characteristics. For unadjusted β values, dimension reduction methods were applied using multidimensional scaling (MDS) [24]. Multidimensional scaling (MDS) is a method to infer the dimensions of the perceptual space of subjects. The results of MDS were visualized in scatter plots. In addition, inter-sample distances were estimated using correlation-based and Manhattan distance and visualized in heat maps.

4.3. Differential methylation analysis

To detect differential methylation between GDM and non-GDM group, the limma (linear models for microarray data) [25] method implemented within the RnBeads package was used [18]. Limma employs empirical Bayesian methods that provide robust results even for small-sized samples. Pair-wise differences in methylation values across all of the CpG sites were estimated after adjusting for confounding factors such as age and sex. The covariate adjustment was done by specifying a design matrix including GDM status, age, and sex. P -values were computed for each of the CpG sites using moderated t -statistics [25]. The linear modeling was done on M -values, which are logistically transformed β values, defined as

$$M = \log_2 \left(\frac{Meth + \alpha}{Unmeth + \alpha} \right)$$

where *Meth* and *Unmeth* are intensities measured by methylated probes and unmethylated probes, respectively, and α , usually set to 1, is an offset value introduced to prevent big changes that occur unexpectedly due to small intensity estimation errors. *M*-value is reported to be more statistically valid for the differential methylation analysis [26]. β values were converted into *M*-values by the following relationship,

$$M = \log_2 \left(\frac{\beta}{1 - \beta} \right).$$

A false-discovery rate (FDR) correction was applied to adjust multiple testing of all the sites [27]. The results of detection for differentially methylated CpGs were visualized in volcano plots with mean methylation differences between two groups and *p*-values across each site presented.

Based on the analysis of differentially methylated CpG sites, differential methylation on region levels, such as CpG islands, genes, and promoters, were detected. Mean of the mean pairwise differences of β values and *p*-values from all sites of the corresponding region were used to detect differentially methylated regions. Combined *p*-values of each genomic region were computed by aggregating *p*-values from correlated significance tests across all sites in the region using weighted inverse chi-square method [28], that is, a generalization of Fisher's inverse chi-square method. A FDR-adjusted *p*-values were also computed to adjust multiple testing of all the regions [27].

5. Function and pathway analysis

For understanding functional relevance with differentially methylated regions, data were further analyzed through the use of QIAGEN's Ingenuity® Pathway Analysis (IPA®, QIAGEN Redwood City, www.qiagen.com/ingenuity), a knowledge-base functional analysis tool. A list of differentially methylated gene regions with p -value < 0.01 and a list of genes associated with differentially methylated promoters with p -value < 0.01 were included respectively for analysis. Statistical significance of overrepresentation of a set of genes in a given canonical pathway, a disease or a biological function was determined using the right-tailed Fisher's exact test.

III. RESULTS

1. Study population

The studied population was grouped into two by *in utero* exposure to maternal GDM. Both groups consisted of 11 males and 8 females. Of the 18 sets of siblings, 10 were sex-concordant, while 8 were sex-discordant. In a GDM-exposed group, ages ranged 4.1-10.26 years old with 9.32 ± 2.40 (mean \pm standard deviation), while ages ranged 4.6-14.44 years old with 5.78 ± 1.33 in the GDM-unexposed group. The between-group mean age differences were 3.54 years old, as the GDM group consisted of 15 first and 4 second children, while the non-GDM group consisted of 1 first, 14 second, 4 third. As two of the sibling pairs had their oldest sibling not included in the study, older siblings of all but one sibling set belong to non-GDM group, while younger siblings of all but one sibling set composed of three siblings belong to GDM group.

Height, weight, and BMI were significantly higher in non-GDM groups than GDM-group due to age differences. To make comparisons of anthropometric measures comparable between the two groups, 2007 Korean National Growth Charts [29] were utilized for estimating the height, weight, and BMI for age and sex. They are presented in z-scores in the table 1. According to 2006 WHO growth standards for preschool children [30] and 2007 WHO growth reference for school-age children and adolescents [31], three in the GDM-exposed group and two in the GDM-unexposed group were categorized as overweight (z-score ≥ 1); one in the GDM-unexposed group were obese (z-score ≥ 2), with her BMI z-score estimated at 2.16; none of the

subjects was classified into thinness (z-score ≤ -2) or severe thinness (z-score ≤ -3). No statistically significant difference of height, weight and BMI in Z-scores were found between groups at the 0.05 significance level. Overall metabolic indicator measures in the absolute scale were higher in the non-GDM group than the GDM-group. However, only waist circumferences and triglyceride levels were significantly different between two groups at p -value < 0.05 .

Table 1. Descriptive characteristics of the study population.

	Non-GDM	GDM	<i>p</i> -value
Age (year)	9.32 ± 2.40 (n = 19)	5.78 ± 1.33 (n = 19)	<0.001
Sex	11 males, 8 females	11 males, 8 females	1
Birth weight (kg)	3.24 ± 0.61 (n = 19)	3.28 ± 0.41 (n = 19)	0.793
Maternal age at delivery (year)	30.37 ± 2.45 (n = 19)	33.95 ± 2.68 (n = 19)	<0.001
Birth order	15 first, 4 second	1 first, 14 second, 4 third	<0.001
Height (cm)	136.11 ± 14.56 (n = 19)	115.26 ± 8.91 (n = 19)	<0.001
Height for age (Z-score)	0.52 ± 0.63 (n = 19)	0.50 ± 0.91 (n = 19)	1
Weight (kg)	34.93 ± 10.13 (n = 19)	21.71 ± 4.29 (n = 19)	<0.001
Weight for age (Z-score)	0.49 ± 0.81 (n = 19)	0.42 ± 0.82 (n = 19)	0.773
BMI (kg/m ²)	18.41 ± 2.31 (n = 19)	16.19 ± 1.16 (n = 19)	0.003
BMI for age (Z-score)	0.34 ± 0.85 (n = 19)	0.21 ± 0.30 (n = 19)	0.603
-2 to -1	1	1	0.753
-1 to 1	15	15	
1 to 2	2	3	
>2	1	0	
Hip circumference (cm)	53.58 ± 2.16 (n = 10)	51.73 ± 1.5 (n = 11)	0.005
Waist circumference (cm)	62.64 ± 7.5 (n = 19)	53.54 ± 4.69 (n = 18)	<0.001

Hip to height ratio	0.42 ± 0.03 (n = 10)	0.46 ± 0.03 (n = 11)	0.008
Waist to height ratio	0.46 ± 0.04 (n = 19)	0.47 ± 0.03 (n = 18)	0.762
Systolic blood pressure (mmHg)	102.25 ± 11.17 (n = 12)	96.67 ± 10.73 (n = 12)	0.195
Diastolic blood pressure (mmHg)	66 ± 12.84 (n = 12)	56.67 ± 8.88 (n = 12)	0.046
Total fat mass (g)	9420.37 ± 4696.08 (n = 19)	4395.74 ± 2074.7 (n = 19)	<0.001
Trunk fat mass (g)	4237.68 ± 2404.61 (n = 19)	1732.05 ± 1029.23 (n = 19)	<0.001
Leg fat mass (g)	3764.32 ± 1715.46 (n = 19)	1858.89 ± 808.29 (n = 19)	<0.001
Total fat mass (%)	27.25 ± 9.47 (n = 19)	20.58 ± 6.18 (n = 19)	0.014
Trunk fat mass (%)	27.02 ± 11.09 (n = 19)	18.51 ± 7.33 (n = 19)	0.009
Leg fat mass (%)	31.07 ± 8.82 (n = 19)	26.39 ± 6.11 (n = 19)	0.065
Whole Body Bone mineral content (g)	1227.21 ± 434.2 (n = 19)	731.29 ± 194.3 (n = 19)	<0.001
Z-score	0.13 ± 0.67 (n = 16)	-0.15 ± 1.13 (n = 13)	0.225
Plasma glucose level (mg/dL)			
Fasting	84.83 ± 14.4 (n = 19)	77.46 ± 12.79 (n = 19)	0.010
0.5-hour	135.89 ± 22.39 (n = 18)	121.28 ± 28.2 (n = 18)	0.084
2-hour	99.86 ± 23.63 (n = 17)	91.02 ± 18.57 (n = 17)	0.087
Plasma insulin level (mg/dL)			
Fasting	14.77 ± 7.14 (n = 18)	13.17 ± 5.49 (n = 18)	0.451
0.5-hour	60.29 ± 35.08 (n = 16)	52.41 ± 43.77 (n = 17)	0.355

2-hour	50.94 ± 49.47 (n = 16)	17.88 ± 12.49 (n = 16)	0.008
Triglyceride (mg/dL)	89.05 ± 46.36 (n = 19)	60.63 ± 23.39 (n = 19)	0.016
High density lipoprotein (mg/dL)	55.63 ± 11.79 (n = 19)	54.95 ± 10.68 (n = 19)	0.673
HbA1C (mmol/L)	5.32 ± 0.17 (n = 19)	5.25 ± 0.29 (n = 19)	0.352

Continuous data were described in mean ± standard deviation and *p*-values were calculated based on Satterthwaite's approximations after fitting a linear mixed model which adjusts for sibling random effects. Categorical data were presented in counts and tested with Pearson's Chi-square test.

2. DNA methylation analysis

2.1. Exploratory analysis

A total of 465,698 CpG sites were included for analysis. Regions of 29,638 genes, 29,795 promoters and 25,851 CpG islands were annotated for the data. The detailed metric description by site or regions is shown in table 2.

Table 2. Summary of DNA methylation data by site/region.

Site/region	Number of sites/regions with coverage across samples	Mean number of sites per region
CpG sites	465,698	-
Genes	29,638	13.1
Promoters	29,795	6.76
CpG islands	25,851	5.6

The methylation levels across all assayed CpG sites showed largely comparable bimodal distributions for GDM and non-GDM groups (Figure 2). According to mean methylation values, for the GDM group, methylation levels of 226,203 CpG sites (48.57%) were hypermethylated ($\beta \geq 0.7$), whereas 173,964 (37.36%) were hypomethylated ($\beta \leq 0.2$) and 53072 sites (14.07%) intermediately methylated ($0.2 < \beta < 0.7$). For non-GDM group, 223,588 sites (48.01%), 172,928 sites (37.13%), and 69,182 sites (14.86%) showed hypermethylation, hypomethylation, and intermediate methylation levels, respectively.

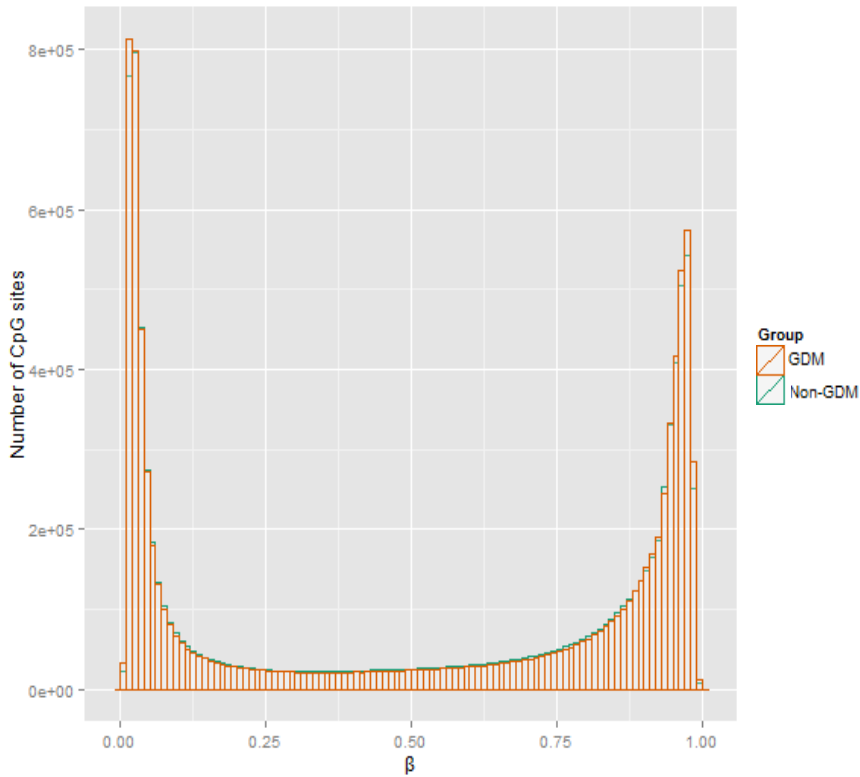


Figure 2. Histogram of β value frequency by the sample group.

Unsupervised hierarchical clustering based on Manhattan distance between samples showed that, of the 19 sibling sets, three of the sibling pairs were closest to each other, and most of the rest of the samples were closely related to each other based on *in utero* GDM exposure status (Figure 3-b). Their multidimensional scaling plot shows moderate distinction between the two groups (Figure 4). Unsupervised hierarchical clustering of methylation values using correlation-based dissimilarity metric generated two distinct clusters, with one cluster containing a single sample (G1-2) and the other containing the

rest of samples (Figure 3-a). The most distinguishing characteristic of the sample, G1-2, was the birth weight of 4.49 kg which was the highest of all subjects. Meanwhile, the sample, N8-1, was most distant to the rest of samples within the cluster. The birth weight of the sample was lowest of all at 2.29 kg. Sibling pairs were closest to each other, overall, except for a sibling pair containing those two samples, G1-2 and N8-1. For a sibling set of N2-1, N2-2 and G2-3, a sibling exposed to impaired glucose tolerance (N2-1) was more related in methylation levels to the other sibling exposed to GDM (G2-3), compared to the sibling exposed to non-diabetic environment (N2-2). For another sibling set of N5-1, G5-2, and G5-3, the siblings exposed to GDM were more closely related to each other, than to the control sibling.

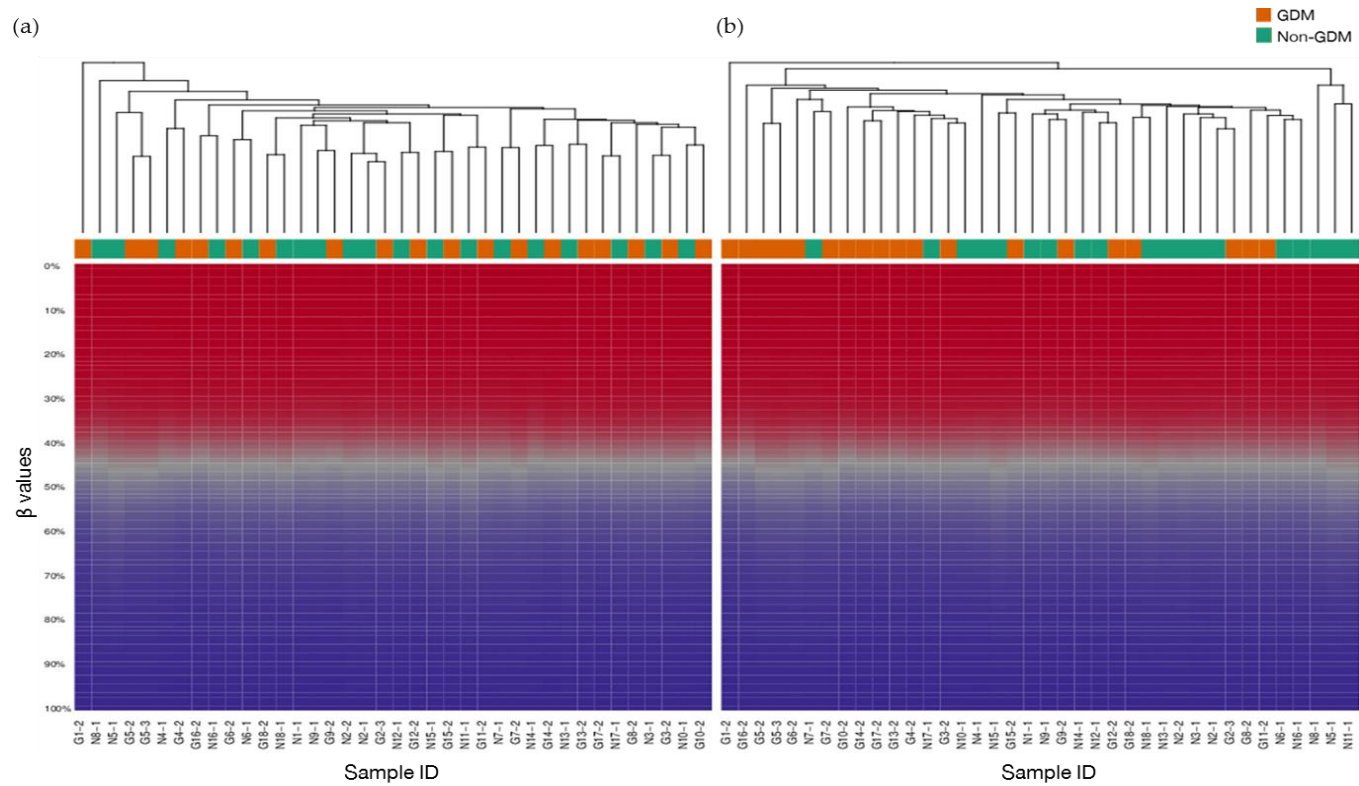


Figure 3. Hierarchical clustering of β values. Inter-sample distances were estimated using (a) correlation-based dissimilarity metric and (b) Manhattan distance. Under the dendrograms are representations of the group to which each sample belongs (GDM group in orange and non-GDM group in green).

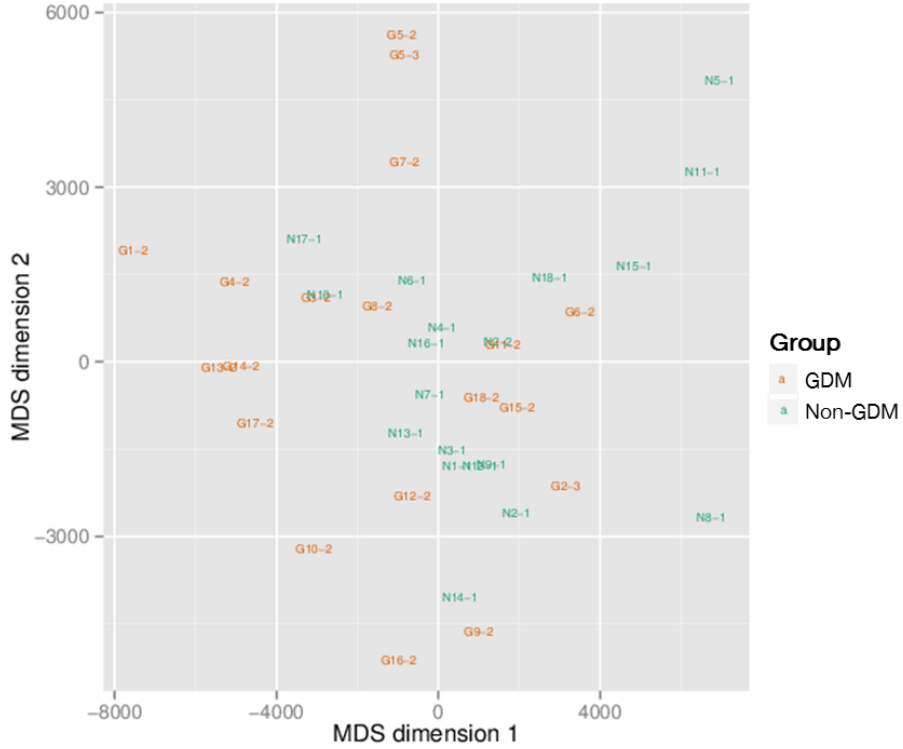


Figure 4. Multidimensional scaling plot of the β values. Each coordinate is denoted by the sample ID, and colored according to *in utero* GDM exposure.

2.2. Differential methylation analysis

Site-specific DNA methylation levels within sibling pairs were compared across all 465,698 CpG sites. For the intra-pair comparisons, two respective sibling sets were trimmed into a pair of siblings. For a sibling set of N2-1, N2-2 and G2-3, a sample exposed to IGT, N2-1, was excluded. For a set of N5-1, G5-2 and G5-3, a sex concordant pair of N2-1 and G2-3 with smaller age gap was selected for the analysis. The two sibling pairs and the rest 16 sibling pairs were included for the pairwise differential methylation analysis.

A total of 18 CpG sites were found to be differentially methylated at a statistical significance of p -values $< 10^{-5}$ (Figure 5, Table 3). No sites, however, were declared to be significant at FDR adjusted p -value < 0.05 for a multiple testing threshold. Of the top 50 CpG sites, 43 are hypermethylated in GDM group compared to non-GDM group, with the mean β value differences ranging from -0.003 to -0.061, while 7 sites are hypomethylated in GDM group, with the mean β value differences from 0.002 to 0.094. For the top ranked CpG sites showing differential methylation at p -values $< 10^{-5}$ ($n = 18$), the pairwise comparison of β values by sibling pair is visualized Figure 6. Of the top 18 sites, 14 are shown to be more hypermethylated in GDM-group compared to non-GDM group. For most of the top sites, highly consistent pairwise differences across sibling pairs are observed.

Differential methylation analyses on the region level of CpG islands, genes and promoters, represented by one or more CpG sites, were performed (Figure 7, and table 4-6). For most of the top CpG island regions, GDM group showed hypomethylation compared to the control group, however, for the top gene and promoter regions, GDM group had relatively more hypermethylation, overall. The region-level analysis showed that there were no regions exhibiting statistical significance at FDR-adjusted p -values < 0.05 . At a combined p -value cut-off of 10^{-3} , 6 CpG islands, 23 genes, and 24 promoters, respectively, were differentially methylated. Most of the top regions of genes and promoters were spanned by one or few CpG sites, while the mean number of CpG sites within genes was 13 and that within promoters was 8.

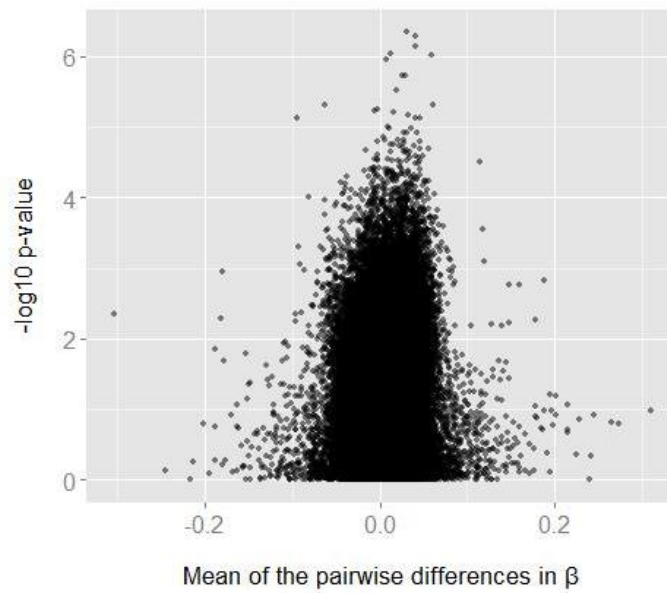


Figure 5. Volcano plots comparing methylation within siblings on site levels.

Table 3. 50 top-ranked differentially methylated CpG sites.

CpG ID	CHR	Start coordinate	Associated genes	Mean β_{GDM}	Mean $\beta_{\text{non-GDM}}$	Mean $\Delta\beta$	p -value	FDR-adjusted p -value
cg01617280	8	8745938	MFHAS1	0.882	0.851	0.031	4.56E-07	0.086
cg13911801	9	35860912	LOC92973	0.862	0.821	0.042	5.06E-07	0.086
cg09304293	6	163475252	PACRG	0.917	0.876	0.04	7.36E-07	0.086
cg08407434	20	43001691	HNF4A	0.957	0.944	0.013	9.11E-07	0.086
cg06928695	17	6384119	PITPNM3	0.781	0.722	0.059	9.44E-07	0.086
cg18255813	6	7195966	RREB1	0.957	0.95	0.007	1.11E-06	0.086
cg10431340	1	161279108	MPZ	0.923	0.894	0.029	1.85E-06	0.109
cg04988367	5	10933006		0.953	0.927	0.026	1.87E-06	0.109
cg06915226	2	70960959	ADD2	0.877	0.858	0.018	3.04E-06	0.157
cg26952618	16	10912545	FAM18A	0.408	0.471	-0.063	4.94E-06	0.193
cg23693245	17	78997300		0.587	0.526	0.061	4.96E-06	0.193
cg02801485	2	202317001	STRADB, TRAK2	0.017	0.02	-0.003	5.66E-06	0.193
cg06319822	16	215960	HBM	0.034	0.039	-0.005	5.98E-06	0.193
cg25558087	17	80346039		0.953	0.937	0.016	6.28E-06	0.193
cg19565171	6	73935200	KHDC1L	0.924	0.892	0.032	6.78E-06	0.193
cg23565826	17	75548591		0.853	0.808	0.045	7.32E-06	0.193

cg09907509	13	37248244	C13orf36	0.116	0.21	-0.094	7.32E-06	0.193
cg07169637	8	103530461		0.847	0.806	0.041	7.46E-06	0.193
cg10408178	14	72219272		0.957	0.948	0.009	1.01E-05	0.223
cg14186245	7	155914047		0.803	0.767	0.036	1.03E-05	0.223
cg10826688	17	43714992	C17orf69, MGC57346	0.922	0.911	0.011	1.06E-05	0.223
cg25233024	19	41066503	SPTBN4	0.909	0.881	0.028	1.19E-05	0.223
cg11506843	8	7153933	FAM90A20	0.774	0.742	0.032	1.19E-05	0.223
cg19009417	10	86004704	RGR	0.882	0.84	0.042	1.20E-05	0.223
cg23910098	11	3187534	OSBPL5	0.982	0.979	0.003	1.42E-05	0.223
cg00637695	5	121759771	SNCAIP	0.899	0.871	0.028	1.51E-05	0.223
cg00871129	20	44176425	SPINLW1	0.904	0.883	0.021	1.53E-05	0.223
cg01359165	15	102033014		0.971	0.964	0.006	1.53E-05	0.223
cg08556772	13	111077142	COL4A2	0.903	0.857	0.046	1.62E-05	0.223
cg25602684	5	37839918	GDNF	0.014	0.017	-0.002	1.64E-05	0.223
cg17761976	12	52491013		0.897	0.87	0.027	1.69E-05	0.223
cg13290745	5	159655866	FABP6	0.944	0.928	0.016	1.76E-05	0.223
cg09483318	17	72130176		0.919	0.906	0.012	1.77E-05	0.223
cg12149666	9	140375580	PNPLA7	0.907	0.868	0.039	1.83E-05	0.223
cg11078990	7	76683209		0.861	0.839	0.022	1.87E-05	0.223
cg11939450	20	18024215	OVOL2	0.899	0.872	0.027	1.91E-05	0.223

cg15083851	7	123080165		0.512	0.456	0.056	2.05E-05	0.223
cg15243478	12	2255779	CACNA1C	0.9	0.854	0.046	2.08E-05	0.223
cg10362591	16	55689865	SLC6A2	0.068	0.084	-0.016	2.12E-05	0.223
cg01671881	20	31804901	C20orf71	0.891	0.86	0.03	2.15E-05	0.223
cg08228856	8	144069849	LOC100133669	0.899	0.87	0.028	2.20E-05	0.223
cg23200572	6	33096501	HLA-DPB2	0.91	0.876	0.034	2.21E-05	0.223
cg26639596	12	124371660	DNAH10	0.893	0.865	0.028	2.21E-05	0.223
cg07386859	16	1872102	HAGH	0.946	0.934	0.012	2.27E-05	0.223
cg20731875	17	14207701	HS3ST3B1, MGC12916	0.948	0.951	-0.003	2.41E-05	0.223
cg16803064	10	131640303	EBF3	0.936	0.904	0.032	2.48E-05	0.223
cg12432161	3	154958565		0.873	0.832	0.041	2.58E-05	0.223
cg14003143	20	42194975	SGK2	0.87	0.822	0.048	2.62E-05	0.223
cg12197579	14	101287676		0.886	0.853	0.033	2.65E-05	0.223
cg11141340	1	26802999	HMGN2	0.798	0.755	0.043	2.76E-05	0.223

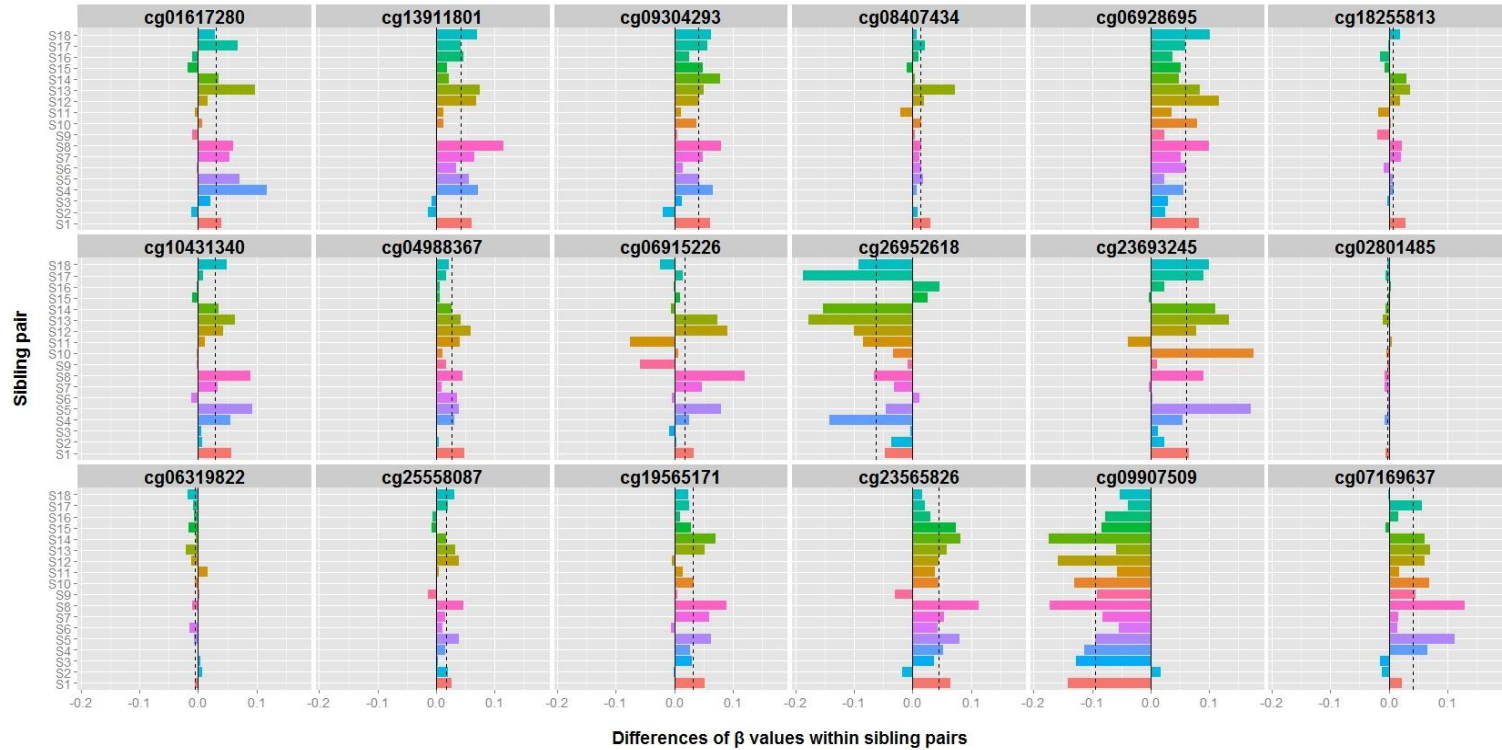


Figure 6. Comparison of differences of β values within sibling pairs. Vertical solid lines and dashed lines represent differences of 0 and the mean differences across samples, respectively. Within-sibling differences were calculated by $\beta_{\text{Absent}} - \beta_{\text{Present}}$ across all samples. The bars were colored according to the sibling pair.

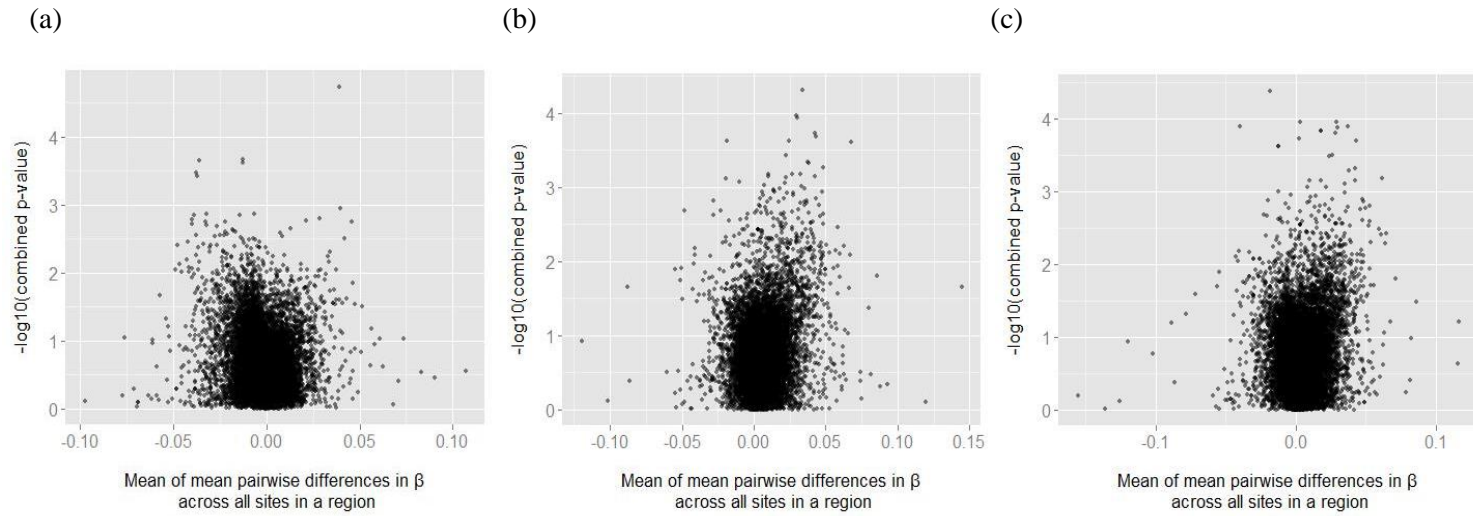


Figure 7. Volcano plots comparing methylation within siblings on region levels. (a) CpG islands, (b) genes, and (c) promoters.

Table 4. 50 top-ranked differentially methylated CpG island regions.

CHR	Start	End	Number of sites associated with the region	Mean β_{GDM}	Mean $\beta_{\text{non-GDM}}$	Mean of the pairwise $\Delta\beta$	Combined p -value	FDR-adjusted combined p -value
9	140375367	140375581	1	0.907	0.868	0.039	1.83E-05	0.474
6	58180155	58180443	1	0.100	0.113	-0.013	2.12E-04	0.674
16	10912160	10912719	6	0.476	0.512	-0.036	2.22E-04	0.674
18	6729694	6730037	2	0.030	0.043	-0.012	2.41E-04	0.674
1	32169538	32169869	4	0.173	0.211	-0.038	3.41E-04	0.674
6	161100093	161100456	3	0.161	0.199	-0.038	3.75E-04	0.674
5	1489800	1490071	3	0.789	0.749	0.040	1.13E-03	0.674
17	1808573	1808773	3	0.336	0.368	-0.032	1.36E-03	0.674
9	34377402	34377610	2	0.059	0.066	-0.007	1.37E-03	0.674
7	2774445	2774655	3	0.173	0.211	-0.039	1.44E-03	0.674
3	184320006	184320218	2	0.131	0.145	-0.014	1.44E-03	0.674
16	55794662	55794939	2	0.572	0.544	0.028	1.61E-03	0.674
5	132155289	132155497	3	0.117	0.157	-0.040	1.67E-03	0.674
9	139397413	139397710	1	0.889	0.868	0.021	1.70E-03	0.674
19	2622523	2622859	1	0.637	0.591	0.046	1.75E-03	0.674
17	17108772	17109701	7	0.121	0.154	-0.033	1.76E-03	0.674
12	54071054	54071265	5	0.057	0.079	-0.022	1.81E-03	0.674

1	112058185	112058590	4	0.133	0.173	-0.040	1.89E-03	0.674
4	682725	683079	2	0.064	0.096	-0.032	2.02E-03	0.674
6	168501903	168502790	4	0.129	0.156	-0.027	2.06E-03	0.674
10	47729706	47729928	1	0.887	0.874	0.013	2.20E-03	0.674
4	206378	206892	4	0.208	0.226	-0.018	2.36E-03	0.674
10	110225928	110226465	3	0.095	0.112	-0.017	2.40E-03	0.674
2	130763484	130763764	1	0.142	0.164	-0.022	2.48E-03	0.674
9	139548206	139548426	1	0.097	0.114	-0.017	2.52E-03	0.674
9	139412168	139412405	1	0.896	0.888	0.007	2.61E-03	0.674
11	94883336	94883565	3	0.366	0.387	-0.021	2.66E-03	0.674
6	21664508	21665178	3	0.114	0.139	-0.025	2.75E-03	0.674
1	209404846	209405473	3	0.349	0.387	-0.038	2.80E-03	0.674
8	29210484	29210801	4	0.115	0.137	-0.023	2.82E-03	0.674
1	175568377	175568808	3	0.121	0.149	-0.028	2.87E-03	0.674
21	46765730	46766088	2	0.868	0.860	0.008	2.90E-03	0.674
9	120507228	120507642	2	0.047	0.061	-0.014	3.06E-03	0.674
19	33726655	33726946	3	0.752	0.710	0.042	3.12E-03	0.674
2	71115928	71116412	3	0.052	0.059	-0.007	3.15E-03	0.674
6	26689804	26690092	2	0.107	0.119	-0.013	3.23E-03	0.674
19	12444033	12444548	4	0.024	0.030	-0.006	3.26E-03	0.674

9	112402768	112403349	2	0.050	0.060	-0.009	3.26E-03	0.674
19	58912261	58913134	3	0.056	0.060	-0.004	3.48E-03	0.674
6	100917206	100917523	4	0.111	0.153	-0.042	3.55E-03	0.674
1	119870227	119870535	3	0.047	0.059	-0.012	3.64E-03	0.674
12	123754050	123754373	3	0.102	0.136	-0.034	3.83E-03	0.674
2	23851841	23852089	3	0.624	0.650	-0.027	3.91E-03	0.674
14	95330984	95331195	2	0.296	0.343	-0.048	3.97E-03	0.674
3	188665276	188665552	3	0.102	0.132	-0.030	3.97E-03	0.674
15	40728263	40728466	2	0.031	0.036	-0.005	4.05E-03	0.674
1	214160799	214161034	2	0.061	0.078	-0.017	4.11E-03	0.674
3	63263990	63264205	2	0.089	0.099	-0.010	4.13E-03	0.674
1	155147186	155147444	2	0.207	0.237	-0.030	4.17E-03	0.674
16	76668836	76669152	2	0.819	0.823	-0.004	4.21E-03	0.674

Table 5. 50 top-ranked differentially methylated gene regions.

Ensemble ID	CHR	Start	End	Associated genes	Number of sites associated with the region	Mean of mean β_{GDM} across all sites in the region	Mean of mean $\beta_{\text{non-GDM}}$ across all sites in the region	Mean of mean $\Delta\beta$ across all sites in the region	Combined p -value	FDR-adjusted combined p -value
ENSG00000197786	11	58125597	58126542	OR5B17	1	0.884	0.850	0.034	4.81E-05	0.559
ENSG00000257845	14	27244701	27291313	LOC101927062	1	0.873	0.844	0.029	1.09E-04	0.559
ENSG00000155833	9	105757593	105780770	CYLC2	1	0.849	0.819	0.030	1.14E-04	0.559
ENSG00000197790	11	4566421	4567374	OR52M1	1	0.752	0.710	0.042	1.89E-04	0.559
ENSG00000231990	9	96619365	96620615	LOC101928014	1	0.893	0.850	0.043	2.06E-04	0.559
ENSG00000224322	7	27401462	27449557		1	0.802	0.778	0.024	2.36E-04	0.559
ENSG00000266441	18	6728820	6729861	LOC101927168	1	0.022	0.041	-0.018	2.41E-04	0.559
ENSG00000241324	7	123977434	123989914	LOC101928211	1	0.799	0.732	0.068	2.44E-04	0.559
ENSG00000261471	16	84627999	84630432		2	0.913	0.891	0.022	3.66E-04	0.559
ENSG00000235408	20	37053843	37053979	SNORA71B	1	0.794	0.757	0.037	4.63E-04	0.559
ENSG00000272192	8	66754987	66755566		1	0.371	0.332	0.038	4.69E-04	0.559
ENSG00000225836	10	91406046	91410579		1	0.885	0.837	0.048	5.42E-04	0.559
ENSG00000226567	3	10801169	10805877	LINC00606	1	0.899	0.867	0.033	6.60E-04	0.559
ENSG00000160862	7	99564343	99573780	AZGP1	3	0.878	0.852	0.027	6.61E-04	0.559
ENSG00000254632	11	76470960	76479267		1	0.967	0.957	0.010	6.68E-04	0.559
ENSG00000249818	4	152826084	152849800	LOC102724700	1	0.794	0.752	0.043	7.07E-04	0.559

ENSG00000235158	3	1049819	1054618		1	0.963	0.953	0.010	7.12E-04	0.559
ENSG00000175676	15	23435096	23448420	GOLGA8EP	1	0.913	0.891	0.022	7.36E-04	0.559
ENSG00000269210	2	39186764	39187483	LOC375196	1	0.895	0.914	-0.020	7.51E-04	0.559
ENSG00000215237	9	14993310	15019727		2	0.053	0.063	-0.010	8.34E-04	0.559
ENSG00000257043	11	18686597	18687095		2	0.899	0.877	0.022	8.84E-04	0.559
ENSG00000207827	6	72113254	72113324	MIR30A	1	0.975	0.968	0.007	9.53E-04	0.559
ENSG00000248492	8	135610314	135612932	ZFAT-AS1	1	0.666	0.632	0.033	9.82E-04	0.559
ENSG00000261060	1	179559507	179560440		1	0.835	0.817	0.018	1.03E-03	0.559
ENSG00000236797	10	77142151	77142602	SPA17P1	1	0.889	0.854	0.035	1.05E-03	0.559
ENSG00000171102	9	136080664	136084630	OBP2B	2	0.920	0.906	0.014	1.08E-03	0.559
ENSG00000205482	7	76682095	76688757		2	0.894	0.880	0.014	1.09E-03	0.559
ENSG00000262343	17	77821857	77823909		1	0.920	0.873	0.048	1.15E-03	0.559
ENSG00000261781	1	18392151	18400906	LOC101927876	1	0.939	0.934	0.006	1.16E-03	0.559
ENSG00000224414	2	218558618	218561486		1	0.958	0.944	0.014	1.17E-03	0.559
ENSG00000267638	17	42193280	42197963		2	0.877	0.850	0.027	1.20E-03	0.559
ENSG00000211581	1	156905923	156906036	MIR765	1	0.881	0.858	0.023	1.21E-03	0.559
ENSG00000261347	11	69282366	69284473		1	0.836	0.829	0.006	1.27E-03	0.559
ENSG00000188069	11	4790209	4791168	OR51F1	1	0.937	0.909	0.027	1.29E-03	0.559
ENSG00000221614	14	101510535	101510620	MIR1185-2	1	0.909	0.887	0.021	1.40E-03	0.559
ENSG00000196248	11	123847368	123848488	OR10S1	1	0.639	0.596	0.043	1.42E-03	0.559

ENSG00000216179	14	101530832	101530915	MIR541	1	0.946	0.935	0.012	1.43E-03	0.559
ENSG00000249928	5	14874509	14874820	UQCRBP3	1	0.977	0.973	0.004	1.44E-03	0.559
ENSG00000204429	9	90795588	90796362		1	0.906	0.869	0.037	1.50E-03	0.559
ENSG00000260253	8	29209937	29210687		4	0.107	0.135	-0.028	1.51E-03	0.559
ENSG00000143556	1	153430220	153433177	S100A7	3	0.668	0.632	0.036	1.55E-03	0.559
ENSG00000226212	7	38380301	38380599	TRGV6	1	0.958	0.941	0.017	1.58E-03	0.559
ENSG00000262836	17	33615	41378		1	0.958	0.949	0.009	1.59E-03	0.559
ENSG00000180988	11	5841544	5842578	OR52N2	1	0.864	0.832	0.032	1.60E-03	0.559
ENSG00000241101	3	64089146	64091732	PRICKLE2-AS2	1	0.819	0.782	0.037	1.63E-03	0.559
ENSG00000231966	1	179798744	179805259		1	0.781	0.736	0.045	1.75E-03	0.559
ENSG00000211582	14	101492357	101492444	MIR758	1	0.974	0.970	0.004	1.75E-03	0.559
ENSG00000223907	1	31984036	31989846	LINC01226	2	0.894	0.859	0.035	1.75E-03	0.559
ENSG00000234953	1	81979565	82023387	LOC101927434	2	0.952	0.948	0.005	1.83E-03	0.559
ENSG00000230043	20	49457152	49457286	TMSB4XP6	1	0.896	0.888	0.008	1.89E-03	0.559

Table 6. 50 top-ranked differentially methylated promoter regions.

Ensemble ID	CHR	Start	End	Associated genes	Number of sites associated with the region	Mean of mean β_{GDM} across all sites in the region	Mean of mean $\beta_{\text{non-GDM}}$ across all sites in the region	Mean of mean $\Delta\beta$ across all sites in the region	Combined p -value	FDR-adjusted combined p -value
ENSG00000250295	8	74268197	74270196	LOC101926926	1	0.860	0.878	-0.018	4.16E-05	0.517
ENSG00000266465	12	58167602	58169601		1	0.958	0.955	0.003	1.12E-04	0.517
ENSG00000221488	1	211383925	211385924		1	0.910	0.881	0.029	1.13E-04	0.517
ENSG00000171532	17	37765531	37767530	NEUROD2	1	0.863	0.826	0.037	1.27E-04	0.517
ENSG00000177447	11	27826490	27828489	CBX3P1	1	0.748	0.788	-0.040	1.31E-04	0.517
ENSG00000200817	11	57793785	57795784	RNU6-899P	1	0.901	0.871	0.030	1.34E-04	0.517
ENSG00000256150	12	2880575	2882574		1	0.950	0.932	0.018	1.51E-04	0.517
ENSG00000255669	12	2879122	2881121	LOC283440	1	0.950	0.932	0.018	1.51E-04	0.517
ENSG00000236083	9	35859481	35861480	OR13E1P	2	0.888	0.859	0.028	1.56E-04	0.517
ENSG00000129473	14	23766499	23768498	BCL2L2	1	0.963	0.961	0.002	1.87E-04	0.552
ENSG00000231990	9	96617865	96619864	LOC101928014	1	0.893	0.850	0.043	2.06E-04	0.552
ENSG00000266441	18	6729362	6731361	LOC101927168	2	0.030	0.043	-0.012	2.41E-04	0.552
ENSG00000088756	18	6728217	6730216	ARHGAP28	2	0.030	0.043	-0.012	2.41E-04	0.552
ENSG00000268473	16	86468991	86470990		1	0.855	0.829	0.026	3.14E-04	0.633
ENSG00000255855	11	94781086	94783085		1	0.876	0.852	0.024	3.25E-04	0.633
ENSG00000236656	1	158464177	158466176		2	0.900	0.857	0.042	4.93E-04	0.633

ENSG00000203258	11	12943973	12945972		1	0.899	0.861	0.038	5.17E-04	0.633
ENSG00000113387	5	32530239	32532238	SUB1	1	0.799	0.737	0.062	6.74E-04	0.633
ENSG00000248991	4	152839120	152841119		1	0.794	0.752	0.043	7.07E-04	0.633
ENSG00000235158	3	1048319	1050318		1	0.963	0.953	0.010	7.12E-04	0.633
ENSG00000268146	1	244226132	244228131		1	0.942	0.924	0.018	7.14E-04	0.633
ENSG00000256143	4	124924	126923		3	0.328	0.292	0.035	7.90E-04	0.633
ENSG00000264458	17	30953257	30955256		1	0.963	0.955	0.008	8.78E-04	0.633
ENSG00000257043	11	18685097	18687096		2	0.899	0.877	0.022	8.84E-04	0.633
ENSG00000252188	7	138309425	138311424		1	0.041	0.053	-0.013	1.01E-03	0.633
ENSG00000236797	10	77142103	77144102	SPA17P1	1	0.889	0.854	0.035	1.05E-03	0.633
ENSG00000176399	9	22445340	22447339	DMRTA1	1	0.028	0.031	-0.004	1.12E-03	0.633
ENSG00000224414	2	218557118	218559117		1	0.958	0.944	0.014	1.17E-03	0.633
ENSG00000240687	2	9777401	9779400		1	0.915	0.904	0.011	1.22E-03	0.633
ENSG00000226428	10	116658981	116660980		1	0.913	0.866	0.047	1.27E-03	0.633
ENSG00000135222	4	70826226	70828225	CSN2	1	0.851	0.825	0.026	1.27E-03	0.633
ENSG00000166676	16	10912152	10914151	TVP23A	9	0.609	0.629	-0.020	1.32E-03	0.633
ENSG00000259348	15	96589660	96591659		1	0.512	0.502	0.010	1.36E-03	0.633
ENSG00000186117	11	55577354	55579353	OR5L1	1	0.796	0.748	0.048	1.37E-03	0.633
ENSG00000259058	14	77390493	77392492		1	0.948	0.940	0.008	1.45E-03	0.633
ENSG00000256879	12	20522697	20524696		2	0.054	0.062	-0.007	1.49E-03	0.633

ENSG00000266121	7	7108036	7110035		1	0.899	0.878	0.021	1.52E-03	0.633
ENSG00000226212	7	38380100	38382099	TRGV6	1	0.958	0.941	0.017	1.58E-03	0.633
ENSG00000250020	5	56813	58812		1	0.807	0.755	0.052	1.60E-03	0.633
ENSG00000168828	9	35869962	35871961	OR13J1	1	0.647	0.610	0.037	1.63E-03	0.633
ENSG00000171243	7	16569706	16571705	SOSTDC1	1	0.929	0.916	0.013	1.66E-03	0.633
ENSG00000234384	13	91144041	91146040	LINC01049	1	0.892	0.842	0.050	1.67E-03	0.633
ENSG00000160200	21	44496554	44498553	CBS	2	0.086	0.091	-0.005	1.72E-03	0.633
ENSG00000197786	11	58126043	58128042	OR5B17	2	0.665	0.619	0.046	1.75E-03	0.633
ENSG00000224876	20	31804374	31806373		2	0.899	0.875	0.024	1.77E-03	0.633
ENSG00000203620	1	32320670	32322669		1	0.856	0.821	0.035	1.86E-03	0.633
ENSG00000119440	9	136103494	136105493	LCN1P1	1	0.953	0.937	0.016	1.88E-03	0.633
ENSG00000254597	8	7624999	7626998	FAM90A10P	2	0.864	0.862	0.003	1.91E-03	0.633
ENSG00000179038	11	66964139	66966138		1	0.728	0.750	-0.023	2.10E-03	0.633
ENSG00000251164	6	8650870	8652869	HULC	1	0.920	0.940	-0.020	2.14E-03	0.633

3. Function and pathway analysis

Pathways and diseases/biological functions associated with differential methylation were identified. For hypermethylated genes and hypermethylated promoter-associated genes, significance of 68 pathways with p -values < 0.05 , equivalent to score $-\log_{10}(p\text{-value}) > 1.30$, were compared with those from hypomethylated genes and hypomethylated promoters-associated genes in Figure 8. *In utero* diabetic exposure-associated diseases and biological functions were also identified and compared with those exhibiting hypomethylation. The diseases or biological functions overrepresented by either hypermethylated gene sets or hypermethylated promoter-associated gene sets (p -values < 0.01) were presented in Figure 9.

The top two canonical pathways associated with hypermethylated genes include ‘Differential Regulation of Cytokine Production in Intestinal Epithelial Cells by IL-17A and IL-17F (p -value = $7.15\text{E-}06$)’ and ‘Crosstalk between Dendritic Cells and Natural Killer Cells (p -value = $2.07\text{E-}05$)’. On the other hand, hypermethylated promoter genes were associated with ‘Phototransduction Pathway (p -value = $8.28\text{E-}05$)’ and ‘Role of Cytokines in Mediating Communication between Immune Cells (p -value = $2.23\text{E-}03$)’.

Cancer (p -values for genes: $1.30\text{E-}02$ - $1.57\text{E-}27$, promoters: $3.71\text{E-}02$ - $3.55\text{E-}31$) was the most significantly associated diseases for both hypermethylated genes and promoter genes. The top molecular and cellular functions, ‘cell-to-cell signaling and interaction (p -values for genes: $1.30\text{E-}02$ - $9.07\text{E-}10$, promoters: $4.07\text{E-}02$ - $1.21\text{E-}05$)’ was shared between hypermethylated genes and promoter-associated genes.

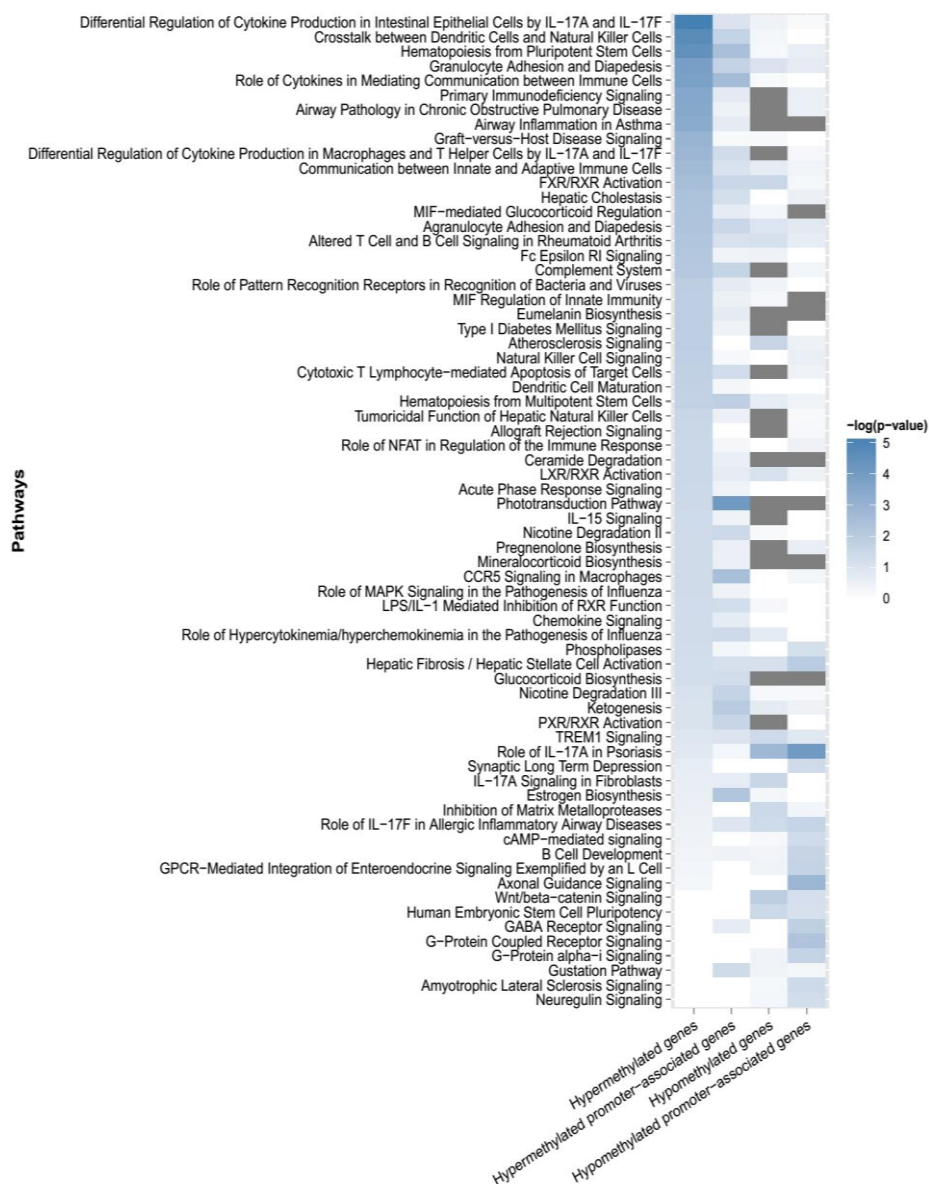


Figure 8. Pathways associated with differentially methylated regions. Pathways significantly associated with hypermethylated regions (p -values < 0.05) were selected and sorted by the significance of hypermethylated genes. The significance levels in $-\log(p\text{-value})$ for association between pathways and gene sets were represented by color. Grey cells represent missing p -values, as no genes are involved in the corresponding pathways.

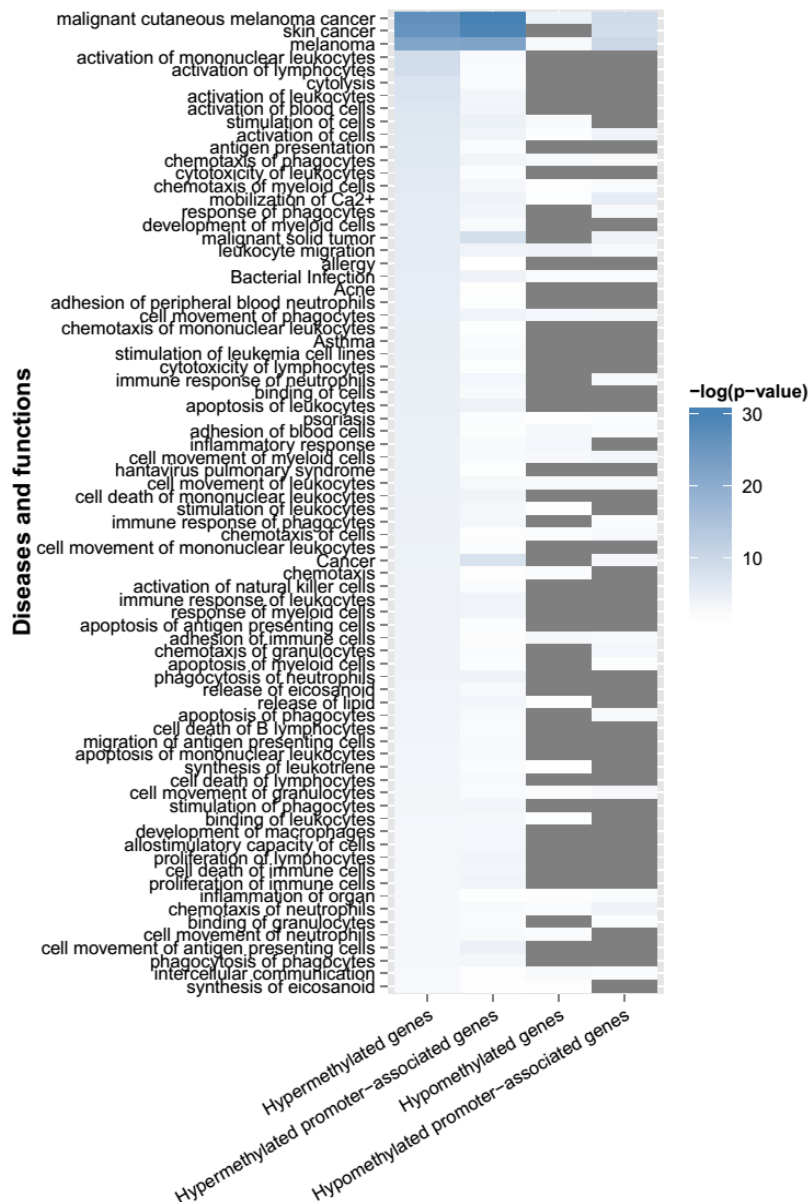


Figure 9. Diseases and biological functions associated with differential methylated regions. Diseases and biological functions significantly associated with hypermethyated regions (p -values < 0.05) were selected and sorted by the significance of hypermethyated genes. The significance levels in $-\log(p\text{-value})$ for association between diseases and biological functions and gene sets were represented by color. Grey cells represent missing p -values, as no genes are involved in the corresponding diseases or functions.

IV. DISCUSSION

To date, several studies have reported differential DNA methylation associated with *in utero* exposures to maternal diabetes including GDM. However, most of the studies adopted a population-based case-control study design that may possibly introduce genetic and environmental confounding effects. This, to the best of our knowledge, is the first study to employ a sibship design discordant for intrauterine exposure to GDM to examine differential DNA methylation status. The sibship study design addressed key confounding issues moderately such as shared genetic factors and familial and shared early-life environmental factors. The discordant sibship study design was further strengthened by intra-sibling pair comparisons made at DNA methylation levels across genome, improving statistical power. Need of pairwise differential methylation analysis was reinforced by the results from the exploratory unsupervised clustering analysis of methylation levels that showed high correlations of methylation profiles within sibling pairs.

Prenatal exposure to diabetic environment was associated with differential methylation at CpG sites and genomic regions including CpG islands, genes and promoters. Among the top differentially methylated CpG sites is cg08407434 located within HNF4A (Hepatic Nuclear Factor 4 Alpha) locus, which is reported to be critical in causing diabetes by reducing the amount of insulin and also involved in maturity-onset diabetes of the young (MODY) [32]. However, differential methylation analysis on gene (number of CpG sites: 20) and promoter (number of CpG sites: 11) levels showed no statistical significance (p -value = 0.16 and 0.14 for genes and promoters,

respectively), even though the mean methylation levels were higher in the GDM group across all sites in HNF4A (mean differences: 0.004 and 0.014 for genes and promoters, respectively). Some of the top differentially methylated regions were non-coding genes, represented by a single CpG probe. AZGP1 (Zinc- α 2-glycoprotein) gene region, hypermethylated in GDM group, is reported to be a tumor suppressor in pancreatic cancer by Kong et al. [33].

As suggested in the analysis on pathways and diseases and biological functions, hypermethylated genes and genes associated with hypermethylated promoters are co-overrepresented. However, not all top pathways associated with hypermethylated promoters were as significant as those associated with hypermethylated genes, and vice versa. It indicates that hypermethylated promoter-specific pathways and biological functions or diseases might exist, and that hypermethylated genes could regulate the pathways while being regulated less by methylation at promoters. Among the top upstream regulators associated with hypermethylation in promoter regions was HNF1A (Hepatic Nuclear Factor 1 Alpha), which is another major gene participating in MODY [34], together with the HNF4A gene.

The major strength of this study is that a discordant sibship design has been employed. However, because advanced maternal age is an established risk factor for GDM [35], younger siblings are more likely to be at higher risk for intrauterine exposures to diabetic environment, resulting in the mean age difference of 3.54 years between groups in this study population as well. Hence, age, as another important confounding factor for DNA methylation status in pediatric populations [36], remains a challenge in this study, even though age

effects were statistically adjusted for. Another limitation of our study was that correlation analysis between DNA methylation levels and gene expression levels was not conducted, as it is critical to examine whether differential methylation leads to differential expression. For a number of CpG sites and regions exhibiting a small difference in their methylation levels made it hard to associate them with biological implications, for example, cg08407434, a CpG site associated with HNF4A loci. And also, cell-type heterogeneity in blood samples, as it is reported in several studies that heterogeneity in cell type proportions is one of the major confounding factors in investigating DNA methylation levels [37]. There also needs to be replication studies examining whether the results are replicated in another population, and, in addition, validation studies that involve a procedure of targeted bisulfate sequencing at differentially methylated regions.

In summary, this is the first epigenome-wide study investigating differential methylation for sibling pairs showing discordance for their intrauterine exposure to maternal diabetes. *In utero* diabetic environmental factors were associated with differential methylation profiles within siblings in their childhood or adolescence, providing supportive evidence that epigenetic signatures, particularly DNA methylation levels, may serve as biomarkers of intrauterine exposure to diabetic environment. Further investigation into pathway and diseases and biological functions associated with differential methylation reinforced the evidence for DOHaD.

V. REFERENCES

1. Schneider, S., et al., *The prevalence of gestational diabetes in advanced economies*. J Perinat Med, 2012. **0**(0): p. 1-10.
2. Jang, H.C., et al., *Pregnancy outcome in Korean women with gestational diabetes mellitus diagnosed by the Carpenter-Coustan criteria*. J Korean Diabetes Assoc, 2004. **28**(2): p. 122-130.
3. Wendland, E.M., et al., *Gestational diabetes and pregnancy outcomes-a systematic review of the World Health Organization (WHO) and the International Association of Diabetes in Pregnancy Study Groups (IADPSG) diagnostic criteria*. BMC Pregnancy Childbirth, 2012. **12**: p. 23.
4. Sobngwi, E., et al., *Effect of a diabetic environment in utero on predisposition to type 2 diabetes*. Lancet, 2003. **361**(9372): p. 1861-5.
5. Dabelea, D., et al., *Intrauterine exposure to diabetes conveys risks for type 2 diabetes and obesity: a study of discordant sibships*. Diabetes, 2000. **49**(12): p. 2208-11.
6. Gillman, M.W., *Developmental origins of health and disease*. N Engl J Med, 2005. **353**(17): p. 1848-50.
7. Heijmans, B.T., et al., *Persistent epigenetic differences associated with prenatal exposure to famine in humans*. Proc Natl Acad Sci U S A, 2008. **105**(44): p. 17046-9.
8. van Abeelen, A.F., et al., *Famine exposure in the young and the risk of type 2 diabetes in adulthood*. Diabetes, 2012. **61**(9): p. 2255-60.
9. El Hajj, N., et al., *Epigenetics and life-long consequences of an adverse nutritional and diabetic intrauterine environment*. Reproduction, 2014. **148**(6): p. R111-R120.
10. Jaenisch, R. and Bird, A., *Epigenetic regulation of gene expression:*

- how the genome integrates intrinsic and environmental signals. *Nat Genet*, 2003. **33 Suppl**: p. 245-54.
11. El Hajj, N., et al., *Metabolic programming of MEST DNA methylation by intrauterine exposure to gestational diabetes mellitus*. *Diabetes*, 2013. **62**(4): p. 1320-8.
 12. del Rosario, M.C., et al., *Potential epigenetic dysregulation of genes associated with MODY and type 2 diabetes in humans exposed to a diabetic intrauterine environment: an analysis of genome-wide DNA methylation*. *Metabolism*, 2014. **63**(5): p. 654-60.
 13. Houde, A.A., et al., *Placental lipoprotein lipase DNA methylation levels are associated with gestational diabetes mellitus and maternal and cord blood lipid profiles*. *J Dev Orig Health Dis*, 2014. **5**(2): p. 132-41.
 14. Lesseur, C., et al., *Maternal obesity and gestational diabetes are associated with placental leptin DNA methylation*. *Am J Obstet Gynecol*, 2014. **211**(6): p. 654 e1-9.
 15. Castillo-Fernandez, J.E., Spector, T.D., and Bell, J.T., *Epigenetics of discordant monozygotic twins: implications for disease*. *Genome Med*, 2014. **6**(7): p. 60.
 16. Kuznetsova, A., Brockhoff, P.B., and Christensen, R.H.B., *lmerTest: Tests in Linear Mixed Effects Models*. 2015.
 17. Team, R.C., *R: A Language and Environment for Statistical Computing*. 2014.
 18. Assenov, Y., et al., *Comprehensive analysis of DNA methylation data with RnBeads*. *Nat Methods*, 2014. **11**(11): p. 1138-40.
 19. Teschendorff, A.E., et al., *A beta-mixture quantile normalization method for correcting probe design bias in Illumina Infinium 450 k DNA methylation data*. *Bioinformatics*, 2013. **29**(2): p. 189-96.
 20. Triche, T.J., Jr., et al., *Low-level processing of Illumina Infinium DNA*

- Methylation BeadArrays*. Nucleic Acids Res, 2013. **41**(7): p. e90.
21. Durinck, S., et al., *Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt*. Nature Protocols, 2009. **4**: p. 1184-1191.
 22. Durinck, S., et al., *BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis*. Bioinformatics, 2005. **21**: p. 3439-3440.
 23. Karolchik, D., et al., *The UCSC Genome Browser Database*. Nucleic Acids Res, 2003. **31**(1): p. 51-4.
 24. Borg, I. and Groenen, P., *Modern multidimensional scaling: theory and applications*. New York: Springer, 1997.
 25. Smyth, G.K., *Linear models and empirical bayes methods for assessing differential expression in microarray experiments*. Stat Appl Genet Mol Biol, 2004. **3**: p. Article3.
 26. Du, P., et al., *Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis*. BMC Bioinformatics, 2010. **11**: p. 587.
 27. Benjamini, Y. and Hochberg, Y., *Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing*. Journal of the Royal Statistical Society Series B-Methodological, 1995. **57**(1): p. 289-300.
 28. Makambi, K.H., *Weighted inverse chi-square method for correlated significance tests*. Journal of Applied Statistics, 2003. **30**(2): p. 225-234.
 29. Moon, J.S., et al., *2007 Korean National Growth Charts: review of developmental process and an outlook*. Korean J Pediatr, 2008. **51**: p. 1-25.
 30. WHO, *The WHO Child Growth Standards*. 2006.
 31. de Onis, M., et al., *Development of a WHO growth reference for school-*

- aged children and adolescents*. Bull World Health Organ, 2007. **85**(9): p. 660-7.
32. Pearson, E.R., et al., *Molecular genetics and phenotypic characteristics of MODY caused by hepatocyte nuclear factor 4alpha mutations in a large European collection*. Diabetologia, 2005. **48**(5): p. 878-85.
 33. Kong, B., et al., *AZGP1 is a tumor suppressor in pancreatic cancer inducing mesenchymal-to-epithelial transdifferentiation by inhibiting TGF-beta-mediated ERK signaling*. Oncogene, 2010. **29**(37): p. 5146-5158.
 34. Horikawa, Y., et al., *Mutation in hepatocyte nuclear factor-1 beta gene (TCF2) associated with MODY*. Nat Genet, 1997. **17**(4): p. 384-5.
 35. Jolly, M., et al., *The risks associated with pregnancy in women aged 35 years or older*. Hum Reprod, 2000. **15**(11): p. 2433-7.
 36. Alisch, R.S., et al., *Age-associated DNA methylation in pediatric populations*. Genome Res, 2012. **22**(4): p. 623-32.
 37. Adalsteinsson, B.T., et al., *Heterogeneity in white blood cells has potential to confound DNA methylation measurements*. PLoS One, 2012. **7**(10): p. e46705.

VI. Abstract in Korean (국문 초록)

어머니의 임신성 당뇨 진단 여부에 따른 자녀들의 DNA 메틸화 수준 비교: 형제자매 내의 비교 연구

서울대학교 보건대학원
보건학과 유전체역학 전공
김 은 애

태아 시절 당뇨 환경에 대한 노출 여부와 대사성 질환 등의 위험의 연관성에 관한 연구들이 보고되고 있으며, 이는 건강과 질병의 발달단계의 기원이라는 가설을 뒷받침하고 있다. 어머니의 당뇨가 자녀들의 건강에 영향을 미치는 것에 관한 기전은 유전자의 발현 조절에 있어 중요한 역할을 하는 DNA 메틸화 패턴의 변화에 의한 것으로 부분적으로 설명되고 있다. 이에 본 연구는 태아 시절 당뇨 환경에 노출 여부와 DNA 메틸화 수준 간의 연관성을 파악하기 위해 수행되었다.

본 연구를 위해 임신성 당뇨 정상 판정을 받고 출산한 경험과 임신성 당뇨 진단 후 출산한 경험이 모두 있는 총 18 명의 여성으로부터 태어난 38 명의 자녀를 모집하였다. 분석을 위해

대상자들의 출생 당시의 자료와 대상자 모집 후 시행하였던 신체 계측 및 생화학적 지표 측정 결과 자료를 수집하였다. 말초 혈액 백혈구 샘플로부터 DNA 를 추출한 다음, Illumina 사의 Infinium HumanMethylation 450 BeadChip 로 어세이하여 유전체 내 총 48,5000 개 이상의 CpG 다이뉴클레오타이드 위치에 대한 DNA 메틸화 수준의 데이터를 얻었다. 샘플 간의 메틸화 프로파일 간의 유사성을 파악하기 위해 기계 학습 분석법의 일종인 자율 학습 기법을 이용한 군집 분석을 수행하였다. 성별 및 연령을 보정한 후 형제자매 내의 DNA 메틸화 수준 차이가 존재하는 CpG 다이뉴클레오타이드 위치를 찾기 위한 분석을 수행하였고, 그 결과를 바탕으로 CpG 섬, 유전자 및 프로모터 수준에서의 DNA 메틸화 차이에 관한 분석을 진행하였다. 나아가 DNA 메틸화 수준의 차이를 보이는 유전자와 연관된 생물학적 경로 및 기능에 관한 분석을 수행하였다.

자율 학습 기법을 통한 군집 분석 및 맨하탄 거리 추정법을 수행한 결과, 대체로 임신성 당뇨 노출 여부에 따라 샘플들이 연관되어 있음을 보였다. 형제 자매 내의 DNA 메틸화 수준의 차이를 비교하기 위한 분석에서는 성별 및 연령을 보정한 후 통계적 유의 수준 10^{-5} 이하에서 총 18 개의 CpG 다이뉴클레오타이드 위치에서 유의한 DNA 메틸화 수준의 차이를 보였으며, 그 중 하나는 HNF4A 유전자좌와 연관이 되어있었다. 통계적 유의 수준 10^{-3} 이하에서 CpG 섬의 경우 총 6 개의 위치에서, 유전자의 경우 총 23 개, 그리고 프로모터의 경우 총 24 개에서 유의한 메틸화 패턴의 차이를 보였다. 관련된 경로 및 기능 분석에서는 높은 메틸화 수준을 보이는 유전자들이 면역 반응에 관여함을 보였다.

본 연구는 어머니의 임신성 당뇨 여부가 불일치한 형제 자매 쌍을 대상으로 수행된 형제자매 내의 전장 후성유전학적 비교 연구이다. 메틸화 수준의 차이를 보이는 유전자를 찾음으로써, 어머니의 임신성 당뇨가 자녀들의 DNA 메틸화 수준의 변화를

일으킨다는 증거를 강화시켰을 뿐만 아니라, 그와 관련된 유전자가 관여하는 생물학적 경로 및 관련 질병 및 기능에 대한 분석을 바탕으로 건강과 질병의 발달단계의 기원의 가설을 뒷받침하였다.

주요어: 임신성 당뇨, DNA 메틸화, 후성유전학, 자궁 내 환경, 건강과 질병 발달단계의 기원 가설, 형제 쌍 연구

학번: 2013-23582